DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems

Marilyn Walker and Rebecca Passonneau
AT&T Shannon Labs
180 Park Ave.
Florham Park, N.J. 07932 {walker,becky}@research.att.com

ABSTRACT

This paper describes a dialogue act tagging scheme developed for the purpose of providing finer-grained quantitative dialogue metrics for comparing and evaluating DARPA COMMUNICATOR spoken dialogue systems. We show that these dialogue act metrics can be used to quantify the amount of effort spent in a dialogue maintaining the channel of communication or, establishing the frame for communication, as opposed to actually carrying out the travel planning task that the system is designed to support. We show that the use of these metrics results in a 7% improvement in the fit in models of user satisfaction. We suggest that dialogue act metrics can ultimately support more focused qualitative analysis of the role of various dialogue strategy parameters, e.g. initiative, across dialogue systems, thus clarifying what development paths might be feasible for enhancing user satisfaction in future versions of these systems.

1. INTRODUCTION

Recent research on dialogue is based on the assumption that dialogue acts provide a useful way of characterizing dialogue behaviors in human-human dialogue, and potentially in human-computer dialogue as well [16, 27, 11, 7, 1]. Several research efforts have explored the use of dialogue act tagging schemes for tasks such as improving recognition performance [27], identifying important parts of a dialogue [12], and as a constraint on nominal expression generation [17]. This paper reports on the development and use of a dialogue act tagging scheme for a rather different task: the evaluation and comparison of spoken dialogue systems in the travel domain. We call this scheme DATE: Dialogue Act Tagging for Evaluation.

Our research on the use of dialogue act tagging for evaluation focuses on the corpus of DARPA COMMUNICATOR dialogues collected in the June 2000 data collection [28]. This corpus consists of 662 dialogues from 72 users calling the nine different COMMUNICATOR travel planning systems. Each system implemented a logfile standard for logging system behaviors and calculating a set of core metrics. Each system utterance and each recognizer result was logged, and user utterances were transcribed and incorporated into the logfiles. The logfile standard supported the calculation of metrics that were hypothesized to potentially affect the user's perception of the system; these included task duration, per turn measures, response latency measures and ASR performance measures. Each dialogue was also hand labelled for task completion.

The hypothesis underlying our approach is that a system's dialogue behaviors have a strong effect on the user's perception of the system. Yet the core metrics that were collected via the logfile standard represent very little about dialogue behaviors. For example, the logging counts system turns and tallies their average length, but doesn't distinguish turns that reprompt the user, or give instructions, from those that present flight information. Furthermore, each COMMUNICATOR system had a unique dialogue strategy and a unique way of achieving particular communicative goals. Thus, in order to explore our hypothesis about the differential effect of these strategies, we needed a way to characterize system dialogue behaviors that would capture such differences yet be applied uniformly to all nine systems. While some sites logged system dialogue behaviors using site-specific dialogue act naming schemes, there existed no scheme that could be applied across sites.

Our goal was thus to develop a dialogue act tagging scheme that would capture important distinctions in this set of dialogues; these distinctions must be useful for testing particular hypotheses about differences among dialogue systems. We also believed that it was important for our tagging scheme to allow for multiple views of each dialogue act. This would allow us, for example, to investigate what part of the task an utterance contributes to separately from what speech act function it serves. A central claim of the paper is that these goals require a tagging scheme that makes distinctions within three orthogonal dimensions of utterance classification: (1) a SPEECH-ACT dimension; (2) a TASK-SUBTASK dimension; and (3) a CONVERSATIONAL-DOMAIN dimension. Figure 1 shows a COMMUNICATOR dialogue with each system utterance classified on these three dimensions. The labels on each utterance are fully described in the remainder of the paper.

Sections 2, 3, and 4, describe the three dimensions of DATE. In these sections, we describe two aspects of our annotation scheme that are not captured in existing tagging schemes, which we believe are important for characterizing how much effort in a dialogue is devoted to the task versus different kinds of dialogue maintenance. Section 5 describes how the dialogue act labels are assigned to system utterances and section 6 discusses results showing that the DATE dialogue act metrics improve models of user satisfaction by an absolute 7% (an increase from 38% to 45%). The dialogueue act metrics that are important predictors of user satisfaction are various kinds of meta-dialogue, apologies and acts that may be landmarks for achieving particular dialogueue subtasks. In section 7 we summarize the paper, discuss our claim that a dialogue annotation

.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2001		**00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **AT&T Labs,Research,180 Park Avenue,Florham Park,NJ,07932**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

scheme is a partial model of a natural class of dialogues, and discuss the ways in which the DATE scheme may be generalizable to other dialogue corpora.

2. CONVERSATIONAL DOMAINS

The CONVERSATIONAL-DOMAIN dimension characterizes each utterance as primarily belonging to one of three arenas of conversational action. The first arena is the domain task, which in this case is air travel booking, and which we refer to below as ABOUT-TASK. The second domain of conversational action is the management of the communication channel, which we refer to as ABOUT-COMMUNICATION. This distinction has been widely adopted [19, 2, 9]. In addition, we identify a third domain of talk that we refer to as ABOUT-SITUATION-FRAME. This domain is particularly relevant for distinguishing human-computer from human-human dialogues, and for distinguishing dialogue strategies across the 9 COMMUNICATOR systems. Each domain is described in this section.

2.1 About-Task

The ABOUT-TASK domain reflects the fact that many utterances in a task-oriented dialogue originate because the goal of the dialogue is to complete a particular task to the satisfaction of both participants. Typically an about-task utterance directly asks for or presents task-related information, or offers a solution to a task goal.

As Figure 1 shows, most utterances are in the ABOUT-TASK dimension, reflecting the fact that the primary goal of the dialogue is to collaborate on the task of making travel arrangements. The task column of Figure 1 specifies the subtask that each task-related utterance contributes to. DATE includes a large inventory of subtasks in the task/subtask dimension in order to make fine-grained distinctions regarding the dialogue effort devoted to the task or its subcomponents. Section 4 will describe the task model in more detail.

2.2 About-Communication

The ABOUT-COMMUNICATION domain reflects the system goal of managing the verbal channel and providing evidence of what has been understood [29, 8, 25]. Although utterances of this type occur in human-human dialogue, they are more frequent in human-computer dialogue, where they are motivated by the need to avoid potentially costly errors arising from imperfect speech recognition. In the COMMUNICATOR corpus, many systems use a conservative strategy of providing feedback indicating the system's understanding of the information provided by the user after each user turn. A typical example is the repetition of the origin and destination cities in Figures 1 and 6. This type of repetition is the IMPLICIT-CONFIRMATION speech-act (see Section 3 below). Some systems used a variable confirmation strategy where some information items may be confirmed as they are understood, but the system requests explicit confirmation of all task parameters before searching the database for matching flights. An example is in Figure 2. Here the system asks for explicit confirmation in SYS3 before going to the database. This is the first opportunity that the user has for making a correction, which he does in USER3. The system then again asks for explicit confirmation of its new understanding, which the user provides in USER4. After the user informs the system that it is a one-way flight in USER6, the system accesses the database. These explicit confirmations have the goal of avoiding a costly database lookup, where the retrieval is conditioned on the wrong parameters.

All implicit and explicit confirmation speech-acts are categorized as ABOUT-COMMUNICATION because they are motivated by the potential errors that the system might make in understanding the caller, or in diagnosing the causes of misunderstandings. In general, any utterance that reflects the system's understanding of something the user said is classified as ABOUT-COMMUNICATION.

Speaker	Utterance
SYS1	....What are your travel plans? ?
USER1	I WISH TO FLY FROM LEXICON KENTUCKY TO SPRINGFIELD MISSOURI THE MORNING OF OCTOBER FIFTH
SYS2	do you have a preferred airline?
USER2	NO
SYS3	I have you going from Kentucky to Gainesville Missouri on October fifth . Is that correct?
USER3	IT'S LEXINGTON KENTUCKY TO SPRINGFIELD MISSOURI
SYS4	I have you going from Lexington to Springfield Missouri on October fifth, Is that correct?
USER4	YES
SYS5	Will you return to Lexington from Springfield?
USER5	NO
SYS6	Is Springfield your final destination?
USER6	YES
SYS7	Hold on while I check availability..... Three options were returned. Option one has a fare of four hundred and thirty three dollars.....

Figure 2: Dialogue Illustrating Variable Confirmation Strategy

A second set of ABOUT-COMMUNICATION utterances are APOLOGIES that the system makes for misunderstandings (see Section 3 below), i.e. utterances such as *I'm sorry. I'm having trouble understanding you.*, or *My mistake again. I didn't catch that.* or *I can see you are having some problems.*

The last category of ABOUT-COMMUNICATION utterances are the OPENINGS/CLOSINGS by which the system greets or says goodbye to the caller. (Again, see Section 3 below.)

2.3 About Situation-Frame

The SITUATION-FRAME domain pertains to the goal of managing the culturally relevant framing expectations. The term is inspired by Goffman's work on the organization and maintenance of social interaction [13, 14]. An obvious example of a framing assumption is that the language of the interaction will be English [13, 14]. Another is that there is an asymmetry between the knowledge and/or agency of the system (or human travel agent) and that of the user (or caller): the user cannot issue an airline ticket.

In developing the DATE tagging scheme, we compared human-human travel planning dialogues collected by CMU with the human-machine dialogues of the June 2000 data collection and noticed a striking difference in the ABOUT-FRAME dimension. Namely, very few ABOUT-FRAME utterances occur in the human-human dialogues, whereas they occur frequently enough in human-computer dialogues that to ignore them is to risk obscuring significant differences in habitability of different systems. In other words, certain differences in dialogue strategies across sites could not be fully represented without such a distinction. Figure 3 provides examples motivating this dimension.

Dialogue acts that are ABOUT-FRAME are cross-classified as one of three types of speech-acts, PRESENT-INFO, INSTRUCTION or APOLOGY. They are not classified as having a value on the TASK-SUBTASK dimension. Most of the ABOUT-FRAME dialogue acts fall into the speech-act category of INSTRUCTIONS, utterances directed at shaping the user's behavior and expectations about how to interact with a machine. Sites differ regarding how much instruction is provided up-front versus within the dialogue; most sites have different utterance strategies for dialogue-initial versus dialogue-

Speech-Act	Example
PRESENT-INFO	*You are logged in as a guest user of A T and T Communicator.*
PRESENT-INFO	*I'll enroll you temporarily as a guest user.*
PRESENT-INFO	*I know about the top 150 cities worldwide.*
PRESENT-INFO	*This call is being recorded for development purposes, and may be shared with other researchers.*
PRESENT-INFO	*I cannot handle rental cars or hotels yet. Please restrict your requests to air travel.*
PRESENT-INFO	*I heard you ask about fares. I can only price an itinerary. I cannot provide information on published fares for individual flights.*
INSTRUCTION	*First, always wait to hear the beep before you say anything*
INSTRUCTION	*You can always start over again completely just by saying: start over.*
INSTRUCTION	*Before we begin, let's go over a few simple instructions.*
INSTRUCTION	*Please remember to speak after the tone. If you get confused at any point you can say start over to cancel your current itinerary.*
APOLOGY	*Sorry, an error has occurred. We'll have to start over.*
APOLOGY	*I am sorry I got confused. Thanks for your patience. Let us try again.*
APOLOGY	*Something is wrong with the flight retrieval.*
APOLOGY	*I have trouble with my script.*

Figure 3: Example About-Frame Utterances

medial instructions. One site gives minimal up-front framing information; further, the same utterances that can occur up-front also occur dialogue-medially. A second site gives no up-front framing information, but it does provide framing information dialogue-medially. Yet a third site gives framing information dialogue-initially, but not dialogue-medially. The remaining sites provide different kinds of general instructions dialogue-initially, e.g. (*Welcome. ...You may say repeat, help me out, start over, or, that's wrong, you can also correct and interrupt the system at any time.*) versus dialogue-medially: (*Try changing your departure dates or times or a nearby city with a larger airport.*) This category also includes statements to the user about the system's capabilities. These occur in response to a specific question or task that the system cannot handle: *I cannot handle rental cars or hotels yet. Please restrict your requests to air travel*. See Figure 3.

Another type of ABOUT-FRAME utterance is the system's attempt to disambiguate the user's utterance; in response to the user specifying *Springfield* as a flight destination, the system indicates that this city name is ambiguous (*I know of three Springfields, in Missouri, Illinois and Ohio. Which one do you want?*). The system's utterance communicates to the user that *Springfield* is ambiguous, and goes further than a human would to clarify that there are only three known options. It is important for evaluation purposes to distinguish the question and the user's response from a simple question-answer sequence establishing a destination. A direct question, such as *What city are you flying to?*, functions as a REQUEST-INFO speech act and solicits information about the task. The context here contrasts with a direct question in that the system has already asked for and understood a response from the caller about the destination city. Here, the function of the system turn is to remediate the caller's assumptions about the frame by indicating the system's confusion about the destination. Note that the question within this pattern could easily be reformulated as a more typical instruction statement, such as *Please specify which Springfield you mean*, or *Please say Missouri, Illinois or Ohio.*.

3. THE SPEECH-ACT DIMENSION

The SPEECH-ACT dimension characterizes the utterance's communicative goal, and is motivated by the need to distinguish the communicative goal of an utterance from its form. As an example, consider the functional category of a REQUEST for information, found in many tagging schemes that annotate speech-acts [24, 18, 6]. Keeping the functional category of a REQUEST separate from the sentence modality distinction between question and statement makes it possible to capture the functional similarity between question and statement forms of requests, e.g., *Can you tell me what time you would like to arrive?* versus *Please tell me what time you would like to arrive.*

In DATE, the speech-act dimension has ten categories. We use familiar speech-act labels, such as OFFER, REQUEST-INFO, PRESENT-INFO, ACKNOWLEDGMENT, and introduce new ones designed to help us capture generalizations about communicative behavior in this domain, on this task, given the range of system and human behavior we see in the data. One new one, for example, is STATUS-REPORT, whose speech-act function and operational definition are discussed below. Examples of each speech-act type are in Figure 4.

Speech-Act	Example
REQUEST-INFO	*And, what city are you flying to?*
PRESENT-INFO	*The airfare for this trip is 390 dollars.*
OFFER	*Would you like me to hold this option?*
ACKNOWLEDGMENT	*I will book this leg.*
STATUS-REPORT	*Accessing the database; this might take a few seconds.*
EXPLICIT-CONFIRM	*You will depart on September 1st. Is that correct?*
IMPLICIT-CONFIRM	*Leaving from Dallas.*
INSTRUCTION	*Try saying a short sentence.*
APOLOGY	*Sorry, I didn't understand that.*
OPENINGS/CLOSINGS	*Hello. Welcome to the C M U Communicator.*

Figure 4: Example Speech Acts

In this domain, the REQUEST-INFO speech-acts are designed to solicit information about the trip the caller wants to book, such as the destination city (*And what city are you flying to?*), the desired dates and times of travel (*What date would you like to travel on*), or information about ground arrangements, such as hotel or car rental (*Will you need a hotel in Chicago?*).

The PRESENT-INFO speech-acts also often pertain directly to the domain task of making travel arrangements: the system presents the user with a choice of itinerary (*There are several flights from Dallas Fort Worth to Salisbury Maryland which depart between eight in the morning and noon on October fifth. You can fly on American departing at eight in the morning or ten thirty two in the morning, or on US Air departing at ten thirty five in the morning.*), as well as a ticket price (*Ticket price is 495 dollars*), or hotel or car options.

OFFERS involve requests by the caller for a system action, such as to pick a flight (*I need you to tell me whether you would like to take this particular flight*) or to confirm a booking (*If this itinerary meets your needs, please press one; otherwise, press zero.*) They typically occur after the prerequisite travel information has been obtained, and choices have been retrieved from the database.

The ACKNOWLEDGMENT speech act characterizes system utterances that follow a caller's acceptance of an OFFER, e.g. *I will book this leg* or *I am making the reservation.*

The STATUS-REPORT speech-act is used to inform the user about the status of the part of the domain task pertaining to the database retrieval, and can include apologies, mollification, requests to be

patient, and so on. Their function is to let the user know what is happening with the database lookup, whether there are problems with it, and what types of problems. While the form of these acts are typically statements, their communication function is different than typical presentations of information; they typically function to keep the user apprised of progress on aspects of the task that the user has no direct information about, e.g. *Accessing the database; this might take a few seconds.* There is also a politeness function to utterances like *Sorry this is taking so long, please hold.*, and they often provide the user with error diagnostics: *The date you specified is too far in advance.*; or *Please be aware that the return date must be later than the departure date.*; or *No records satisfy your request.*; or *There don't seem to be any flights from Boston.*

The speech-act inventory also includes two types of speech acts whose function is to confirm information that has already been provided by the caller. In order to identify and confirm the parameters of the trip, systems may ask the caller direct questions, as in SYS3 and SYS4 in Figure 2. These EXPLICIT-CONFIRM speech acts are sometimes triggered by the system's belief that a misunderstanding may have occurred. A typical example is *Are you traveling to Dallas?*. An alternative form of the same EXPLICIT-CONFIRM speech-act type asserts the information the system has understood and asks for confirmation in an immediately following question: *I have you arriving in Dallas. Is that correct?* In both cases, the caller is intended to provide a response.

A less intrusive form of confirmation, which we tag as IMPLICIT-CONFIRM, typically presents the user with the system's understanding of one travel parameter immediately before asking about the next parameter. Depending on the site, implicit information can either precede the new request for information, as in *Flying to Tokyo. What day are you leaving?*, or can occur within the same utterance, as in *What day do you want to leave London?* More rarely, an implicit confirmation is followed by PRESENT-INFO: *a flight on Monday September 25. Delta has a flight departing Atlanta at nine thirty.* One question about the use of implicit confirmation strategy is whether the caller realizes they can correct the system when necessary [10]. Although IMPLICIT-CONFIRMS typically occur as part of a successful sequence of extracting trip information from the caller, they can also occur in situations where the system is having trouble understanding the caller. In this case, the system may attempt to instruct the user on what it is doing to remediate the problem in between an IMPLICIT-CONFIRM and a REQUEST-INFO: *So far, I have you going from Tokyo. I am trying to assemble enough information to pick a flight. Right now I need you to tell me your destination.*

We have observed that INSTRUCTIONS are a speech-act type that distinguishes these human-computer travel planning dialogues from corresponding human-human travel planning dialogues. Instructions sometimes take the form of a statement or an imperative, and are characterized by their functional goal of clarifying the system's own actions, correcting the user's expectations, or changing the user's future manner of interacting with the system. Dialogue systems are less able to diagnose a communication problem than human travel agents, and callers are less familiar with the capabilities of such systems. As noted above, some systems resort to explicit instructions about what the system is doing or is able to do, or about what the user should try in order to assist the system: *Try asking for flights between two major cities*; or *You can cancel the San Antonio, Texas, to Tampa, Florida flight request or change it. To change it, you can simply give new information such as a new departure time.* Note that INSTRUCTIONS, unlike the preceding dialogue act types, do not directly involve a domain task.

Like the INSTRUCTION speech-acts, APOLOGIES do not address a domain task. They typically occur when the system encounters problems, for example, in understanding the caller (*I'm sorry, I'm having trouble understanding you*), in accessing the database (*Something is wrong with the flight retrieval*), or with the connection (*Sorry, we seem to have a bad connection. Can you please call me back later?*).

The OPENING/CLOSING speech act category characterizes utterances that open and close the dialogue, such as greetings or goodbyes [26]. Most of the dialogue systems open the interactions with some sort of greeting—*Hello, welcome to our Communicator flight travel system,* and end with a sign-off or salutation—*Thank you very much for calling. This session is now over.* We distinguish these utterances from other dialogue acts, but we do not tag openings separate from closings because they have a similar function, and can be distinguished by their position in the discourse. We also include in this category utterances in which the systems survey the caller as to whether s/he got the information s/he needed or was happy with the system.

4. THE TASK-SUBTASK DIMENSION

The TASK-SUBTASK dimension refers to a task model of the domain task that the system is designed to support and captures distinctions among dialogue acts that reflect the task structure.[1] Our domain is air travel reservations, thus the main communicative task is to specify information pertaining to an air travel reservation, such as the destination city. Once a flight has been booked, ancillary tasks such as arranging for lodging or a rental car become relevant. The fundamental motivation for the TASK-SUBTASK dimension in the DATE scheme is to derive metrics related to subtasks in order to quantify how much effort a system expends on particular subtasks.[2]

This dimension distinguishes among 13 subtasks, some of which can also be grouped at a level below the top level task. The subtasks and examples are in Figure 5. The TOP-LEVEL-TRIP task describes the task which contains as its subtasks the ORIGIN, DESTINATION, DATE, TIME, AIRLINE, TRIP-TYPE, RETRIEVAL and ITINERARY tasks. The GROUND task includes both the HOTEL and CAR subtasks.

Typically each COMMUNICATOR dialogue system acts as though it utilizes a task model, in that it has a particular sequence in which it will ask for task information if the user doesn't take the initiative to volunteer this information. For example, most systems ask first for the origin and destination cities, then for the date and time. Some systems ask about airline preference and others leave it to the caller to volunteer this information. A typical sequence of tasks for the flight planning portion of the dialogue is illustrated in Figure 6.

As Figure 6 illustrates, any subtask can involve multiple speech acts. For example, the DATE subtask can consist of acts requesting, or implicitly or explicitly confirming the date. A similar example is provided by the subtasks of CAR (rental) and HOTEL, which include dialogue acts requesting, confirming or acknowledging arrangements to rent a car or book a hotel room on the same trip.

[1] This dimension is used as an elaboration of each speech-act type in other tagging schemes [24].

[2] It is tempting to also consider this dimension as a means of inferring discourse structure on the basis of utterance level labels, since it is widely believed that models of task structure drive the behavior of dialogue systems [23, 3, 22], and the relationship between discourse structure and task structure has been a core topic of research since Grosz's thesis [15]. However, we leave the inference of discourse structure as a topic for future work because the multifunctionality of many utterances suggests that the correspondence between task structure and dialogue structure may not be as straightforward as has been proposed in Grosz's work [30].

Task	Example
TOP-LEVEL-TRIP	*What are your travel plans?*
ORIGIN	*And, what city are you leaving from?*
DESTINATION	*And, where are you flying to?*
DATE	*What day would you like to leave?*
TIME	*Departing at what time?.*
AIRLINE	*Did you have an airline preference?*
TRIP-TYPE	*Will you return to Boston from San Jose?*
RETRIEVAL	*Accessing the database; this might take a few seconds.*
ITINERARY	*The airfare for this trip is 390 dollars.*
GROUND	*Did you need to make any ground arrangements?.*
HOTEL	*Would you like a hotel near downtown or near the airport?.*
CAR	*Do you need a car in San Jose?*

Figure 5: Example Utterances for each Subtask

Speaker	Utterance	Task
SYS	Welcome .... What trip would you like to take?	TOP-LEVEL-TRIP
USER	STARTING FROM ATLANTA GEORGIA TO LONDON ENGLAND	
SYS	flying from atlanta to london.	ORIGIN, DESTINATION
SYS	What day are you leaving Atlanta?	DATE
USER	MONDAY SEPTEMBER TWENTY FIVE	
SYS	on monday, september twenty fifth.	DATE
SYS	About what time do you want to leave?	TIME
USER	[UH] DAYTIME	
SYS	Leaving in the daytime.	TIME
SYS	Please stand by while I look up the flight schedules on the web.	RETRIEVAL
SYS	OK, got them. I have 13 flights. The first flight is on American at six fifty nine eh M, arriving at ten forty five PM, with a connection in Chicago. . Is that OK?	ITINERARY
USER	[UH] YES OKAY	
SYS	Do you want a return flight from London to Atlanta?	TRIP-TYPE
USER	[UH] YES	
SYS	Returning from london to atlanta.	TRIP-TYPE
SYS	What day are you leaving London?	DATE
USER	[UH] LEAVING [UM] MONDAY OCTOBER THE SECOND	
SYS	on monday, october second.	DATE
SYS	About what time do you want to leave?	TIME
USER	ANY TIME'S OKAY	
SYS	Please stand by while I look up the flight schedules on the web... OK, got them.	RETRIEVAL, ITINERARY

Figure 6: Dialogue Illustrating a Typical Task Sequence

There are also differences in how each site's dialogue strategy reflects it conceptualization of the travel planning task. For example, some systems ask the user explicitly for their airline preferences whereas others do not (the systems illustrated in Figures 1 and 6 do not, wherase the one in Figure 2 does). Another difference is whether the system asks the user explicitly whether s/he wants a round-trip ticket. Some systems ask this information early on, and search for both the outbound and the return flights at the same time. Other systems do not separately model round-trip and multi-leg trips. Instead they ask the user for information leg by leg, and after requesting the user to select an itinerary for one leg of the flight, they ask whether the user has an additional destination. A final difference was that, in the June 2000 data collection, some systems such as the one illustrated in Figure 1 included the ground arrangements subtasks, and others did not.

5. IMPLEMENTATION

Our focus in this work is in labelling the system side of the dialogue; our goal was to develop a fully automatic 100% correct dialogue parser for the limited range of utterances produced by the 9 COMMUNICATOR systems. While we believe that it would be useful to be able to assign dialogue acts to both sides of the conversation, we expect that to require hand-labelling [1]. We also believe that in many cases the system behaviors are highly correlated with the user behaviors of interest; for example when a user has to repeat himself because of a misunderstanding, the system has probably prompted the user multiple times for the same item of information and has probably apologized for doing so. Thus this aspect of the dialogue would also be likely to be captured by the APOLOGY dialogue act and by counts of effort expended on the particular subtask.

We implemented a pattern matcher that labels the system side of each dialogue. An utterance or utterance sequence is identifed automatically from a database of patterns that correspond to the dialogue act classification we arrived at in cooperation with the site developers. Where it simplifies the structure of the dialogue parser, we assign two adjacent utterances that are directed at the same goal the same DATE label, thus ignoring the utterance level segmentation, but we count the number of characters used in each act. Since some utterances are generated via recursive or iterative routines, some patterns involve wildcards.

The current implementation labels the utterances with tags that are independent of any particular markup-language or representation format. We have written a transducer that takes the labelled dialogues and produces HTML output for the purpose of visualizing the distribution of dialogue acts and meta-categories in the dialogues. An additional summarizer program is used to produce a summary of the percentages and counts of each dialogue act as well as counts of meta-level groupings of the acts related to the different dimensions of the tagging scheme. We intend to use our current representation to generate ATLAS compliant representations [4].

6. RESULTS

Our primary goal was to achieve a better understanding of the qualitative aspects of each system's dialogue behavior. We can quantify the extent to which the dialogue act metrics have the potential to improve our understanding by applying the PARADISE framework to develop a model of user satisfaction and then examining the extent to which the dialogue act metrics improve these models [31]. In other work, we show that given the standard metrics collected for the COMMUNICATOR dialogue systems, the best model accounts for 38% of the variance in user satisfaction [28].

When we retrain these models with the dialogue act metrics extracted by our dialogue parser, we find that many metrics are significant predictors of user satisfaction, and that the model fit increases from 38% to 45%. When we examine which dialogue metrics are significant, we find that they include several types of meta-dialogue such as explicit and implicit confirmations of what the user said, and acknowledgments that the system is going to go ahead and do the action that the user has requested. Significant negative predictors include apologies. On interpretation of many of the significant predictors is that they are landmarks in the dialogue for achievement of particular subtasks. However the predictors based on the core metrics included a ternary task completion metric that captures

succinctly whether any task was achieved or not, and whether the exact task that the user was attempting to accomplish was achieved. A plausible explanation for the increase in the model fits is that user satisfaction is sensitive to exactly how far through the task the user got, even when the user did not in fact complete the task. The role of the other significant dialogue metrics are plausibly interpreted as acts important for error minimization. As with the task-related dialogue metrics, there were already metrics related to ASR performance in the core set of metrics. However, several of the important metrics count explicit confirmations, one of the desired date of travel, and the other of all information before searching the database, as in utterances SYS3 and SYS4 in Figure 2.

7. DISCUSSION

This paper has presented DATE, a dialogue act tagging scheme developed explicitly for the purpose of comparing and evaluating spoken dialogue systems. We have argued that such a scheme needs to make three important distinctions in system dialogue behaviors and we are investigating the degree to which any given type of dialogue act belongs in a single category or in multiple categories.

We also propose the view that a tagging scheme be viewed as a partial model of a natural class of dialogues. It is a model to the degree that it represents claims about what features of the dialogue are important and are sufficiently well understood to be operationally defined. It is partial in that the distributions of the features and their relationship to one another, i.e., their possible manifestations in dialogues within the class, are an empirical question.

The view that a dialogue tagging scheme is a partial model of a class of dialogues implies that a pre-existing tagging scheme can be re-used on a different research project, or by different researchers, only to the degree that it models the same natural class with respect to similar research questions, is sufficient for expressing observations about what actually occurs within the current dialogues of interest, and is sufficiently well-defined that high reliability within and across research sites can be achieved. Thus, our need to modify existing schemes was motivated precisely to the degree that existing schemes fall short of these requirements. Other researchers who began with the goal of re-utilizing existing tagging schemes have also found it necessary to modify these schemes for their research purposes [11, 18, 7].

The most substantial difference between our dialogue act tagging scheme and others that have been proposed is in our expansion of the two-way distinction between dialogue *tout simple* vs. meta-dialogue, into a three-way distinction among the immediate dialogue goals, meta-dialogue utterances, and meta-situation utterances. Depending on further investigation, we might decide these three dimensions have equal status within the overall tagging scheme (or within the overall dialogue-modeling enterprise), or that there are two types of meta-dialogue: utterances devoted to maintaining the channel, versus utterances devoted to establishing/maintaining the frame. Further, in accord with our view that a tagging scheme is a partial model, and that it is therefore necessarily evolving as our understanding of dialogue evolves, we also believe that our formulation of any one dimension, such as the speech-act dimension, will necessarily differ from other schemes that model a speech-act dimension.

Furthermore, because human-computer dialogue is at an early stage of development, any such tagging scheme must be a moving target, i.e., the more progress is made, the more likely it is we may need to modify along the way the exact features used in an annotation scheme to characterize what is going on. In particular, as system capabilities become more advanced in the travel domain, it will probably be necessary to elaborate the task model to capture different aspects of the system's problem solving activities. For example, our task model does not currently distinguish between different aspects of information about an itinerary, e.g. between presentation of price information and presentation of schedule information.

We also expect that some domain-independent modifications are likely to be necessary as dialogue systems become more successful, for example to address the dimension of "face", i.e. the positive politeness that a system shows to the user [5]. As an example, consider the difference between the interpretation of the utterance, *There are no flights from Boston to Boston*, when produced by a system vs. when produced by a human travel agent. If a human said this, it would be be interpretable by the recipient as an insult to their intelligence. However when produced by a system, it functions to identify the source of the misunderstanding. Another distinction that we don't currently make which might be useful is between the initial presentation of an item of information and its re-presentation in a summary. Summaries arguably have a different communicative function [29, 7]. Another aspect of function our representation doesn't capture is rhetorical relations between speech acts [20, 21].

While we developed DATE to answer particular research questions in the COMMUNICATOR dialogues, there are likely to be aspects of DATE that can be applied elsewhere. The task dimension tagset reflects our model of the domain task. The utility of a task model may be general across domains and for this particular domain, the categories we employ are presumably typical of travel tasks and so, may be relatively portable.

The speech act dimension includes categories typically found in other classifications of speech acts, such as REQUEST-INFO, OFFER, and PRESENT-INFO. We distinguish information presented to the user about the task, PRESENT-INFO, from information provided to change the user's behavior, INSTRUCTION, and from information presented in explanation or apology for an apparent interruption in the dialogue, STATUS-REPORT. The latter has some of the flavor of APOLOGIES, which have an inter-personal function, along with OPENINGS/CLOSINGS. We group GREETINGS and SIGN-OFFS into the single category of OPENINGS/CLOSINGS on the assumption that politeness forms make less contribution to perceived system success than the system's ability to carry out the task, to correct misunderstandings, and to coach the user.

Our third dimension, conversational-domain, adds a new category, ABOUT-SITUATION-FRAME, to the more familiar distinction between utterances directed at a task goal vs. utterances directed at a maintaining the communication. This distinction supports the separate classification of utterances directed at managing the user's assumptions about how to interact with the system on the air travel task. As we mention above, the ABOUT-SITUATION-FRAME utterances that we find in the human-computer dialogues typically did not occur in human-human air travel dialogues. In addition, as we note above, one obvious difference in the dialogue strategies implemented at different sites had to do with whether these utterances occurred upfront, within the dialogue, or both.

In order to demonstrate the utility of dialogue act tags as metrics for spoken dialogue systems, we show that the use of these metrics in the application of PARADISE [31] improves our model of user satisfaction by an absolute 7%, from 38% to 45%. This is a large increase, and the fit of these models are very good for models of human behavior. We believe that we have only begun to discover the ways in which the output of the dialogue parser can be used. In future work we will examine whether other representations derived from the metrics we have applied, such as sequences or structural relations between various types of acts might improve our performance model further. We are also collaborating with other mem-

bers of the COMMUNICATOR community who are investigating the use of dialogue act and initiative tagging schemes for the purpose of comparing human-human to human-computer dialogues [1].

8. ACKNOWLEDGMENTS

This work was supported under DARPA GRANT MDA 972 99 3 0003 to AT&T Labs Research. Thanks to Payal Prabhu and Sungbok Lee for their assistance with the implementation of the dialogue parser. We also appreciate the contribution of J. Aberdeen, E. Bratt, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, A. Rudnicky, S. Seneff, and D. Stallard who helped us understand how the DATE classification scheme applied to their COMMUNICATOR systems' dialogues.

9. REFERENCES

[1] J. Aberdeen and C. Doran. Human-computer and human-human dialogues. DARPA Communicator Principle Investigators Meeting (Philadelphia, PA USA). http://www.dsic-web.net/ito/meetings/communicator sep2000/, September, 2000.

[2] J. Allen and M. Core. Draft of DAMSL: Dialog act markup in several layers. Coding scheme developed by the MultiParty group, 1st Discourse Tagging Workshop, University of Pennsylvania, March 1996, 1997.

[3] J. F. Allen. Recognizing intentions from natural language utterances. In M. Brady and R. Berwick, editors, *Computational Models of Discourse*. MIT Press, 1983.

[4] S. Bird and M. Liberman. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60, 2001.

[5] P. Brown and S. Levinson. *Politeness: Some universals in language usage*. Cambridge University Press, 1987.

[6] J. C. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23-1:13–33, 1997.

[7] R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, and C. Soria. Building a corpus of annotated dialogues: the ADAM experience. In *Proc. of the Conference Corpus-Linguistics-2001, Lancaster, U.K.*, 2001.

[8] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[9] S. L. Condon and C. G. Cech. Functional comparison of face-to-face and computer-mediated decision-making interactions. In S. Herring, editor, *Computer-Mediated Converstaion*. John Benjamins, 1995.

[10] M. Danieli and E. Gerbino. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, 1995.

[11] B. Di Eugenio, P. W. Jordan, J. D. Moore, and R. H. Thomason. An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proceedings of the Thirty-sixth Conference of the Association for Computational Linguistics*, 1998.

[12] M. Finke, M. Lapata, A. Lavie, L. Levin, L. M. Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. Clarity: Inferring discourse structure from speech. In *American Association for Artificial Intelligence (AAAI) Symposium on Applying Machine Learning to Discourse Processing Proceedings, Stanford, California*, March 1998.

[13] E. Goffman. *Frame Analysis: An Essay on the Organization of Experience*. Harper and Row, New York, 1974.

[14] E. Goffman. *Forms of Talk*. University of Pennsylvania Press, Philadelphia, Pennsylvania, USA, 1981.

[15] B. J. Grosz. The representation and use of focus in dialogue understanding. Technical Report 151, SRI International, 333 Ravenswood Ave, Menlo Park, Ca. 94025, 1977.

[16] A. Isard and J. C. Carletta. Replicability of transaction and action coding in the map task corpus. In M. Walker and J. Moore, editors, *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pages 60–67, 1995.

[17] P. W. Jordan. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. PhD thesis, Intelligent Systems Program, University of Pittsburgh, 2000.

[18] D. Jurafsky, E. Shriberg, and D. Biasca. Swbd-damsl labeling project coder's manual. Technical report, University of Colorado, 1997. available as http://stripe.colorado.edu/ jurafsky/manual.august1.html.

[19] D. Litman. Plan recognition and discourse analysis: An integrated approach for understanding dialogues. Technical Report 170, University of Rochester, 1985.

[20] D. Marcu. Perlocutions: The achilles' heel of speech act theory. *Journal of Pragmatics*, 1999.

[21] M. G. Moser, J. Moore, and E. Glendening. Instructions for coding explanations: Identifying segments, relations and minimal units. Technical Report 96-17, University of Pittsburgh, Department of Computer Science, 1996.

[22] R. Perrault and J. Allen. A plan-based analysis of indirect speech acts. *American Journal of Computational Linguistics*, 6:167–182, 1980.

[23] R. Power. *A Computer Model of Conversation*. PhD thesis, University of Edinburgh, 1974.

[24] N. Reithinger and E. Maier. Utilizing statistical speech act processing in verbmobil. In *ACL 95*, 1995.

[25] D. R.Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3):575–599, 1992.

[26] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8:289–327, 1977.

[27] E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech: Special Issue on Prosody and Conversation*, 2000.

[28] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Submitted to EUROSPEECH 2001*, 2001.

[29] M. A. Walker. Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pages 345–351, 1992.

[30] M. A. Walker. Limited attention and discourse structure. *Computational Linguistics*, 22-2:255–264, 1996.

[31] M. A. Walker, C. A. Kamm, and D. J. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

Speaker	Utterance	Speech Act	Task	C-Domain
SYS	Welcome.	opening/closing	nil	task
	You are logged in as a guest user.	instruct	nil	frame
	You may say repeat, help me out, start over, or, that's wrong,	instruct	nil	frame
	you can also correct and interrupt the system at any time.	instruct	nil	frame
	What airport woodja like to fly out of?	request-info	origin	task
USER	MIAMI FLORIDA			
SYS	Leaving from Miami,	implicit-conf	origin	communication
	And, what city are you flying to?	request-info	dest	task
USER	MINNEAPOLIS MINNESOTA			
SYS	Flying from Miami to Minneapolis,	implicit-conf	origin, dest	communication
	What date would you like to fly?	request-info	date	task
USER	UH MONDAY OCTOBER TWENTY THIRD			
SYS	Leaving on the 23rd of October,	implicit-conf	date	communication
	And what time didja wanna leave?	request-info	time	task
USER	UH LATE MORNING			
SYS	From Minneapolis.	implicit-conf	origin	communication
	Are you returning to Miami?	request-info	trip-type	task
USER	UM			
SYS	Are you going back to Miami?	request-info	trip-type	task
	Please say yes or no.	instruct	nil	frame
USER	NO			
SYS	Okay, it will just take a moment while I look.	status-report	retrieval	task
SYS	I found 3 flights from Miami to Minneapolis on October 23rd . For option one I have a flight with United departing at 10 A M and arriving at 3 P M with one stopover and	present-info	itinerary	task
	Ticket price is 390 dollars.	present-info	price	task
	Please say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	NEXT OPTION			
SYS	For option two I have a flight with Delta departing at 12 and arriving at 4 P M non-stop.	present-info	itinerary	task
	Ticket price is 450 dollars.	present-info	itinerary	task
	Would you like me to hold this option?	offer	itinerary	task
USER	NEXT OPTION			
SYS	For the last option I have a flight with Northwest departing at 2 P M and arriving at 7 P M with 1 stopover and	present-info	itinerary	task
	Ticket price is 470 dollars.	present-info	itinerary	task
	say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	NEXT OPTION			
SYS	Here is your third option again. a flight with Northwest departing at 2 P M and arriving at 7 P M with 1 stopover and	present-info	itinerary	task
	Ticket price is 470 dollars.	present-info	itinerary	task
	say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	I'LL TAKE IT			
SYS	Great! I am adding this flight to your itinerary.	acknowledge	booking	task
	Okay, Is Minneapolis your final destination?	request-info	trip-type	task
USER	YES IT IS			
SYS	Would you like to make any ground arrangements?	offer	ground	task
USER	NO			
SYS	Now, would you like to hear a summary of your itinerary?	offer	itinerary	task
USER	YES PLEASE			
SYS	Okay, Here's your itinerary. On October 23rd Northwest Flight 123 departs Miami at 2 P M, arrives Pittsburgh at 4 30 P M connecting to Northwest Flight 146 that departs Pittsburgh at 5 P M and arrives Minneapolis at 7 P M.	present-info	itinerary	task

Figure 1: Dialogue Illustrating the Speech Act, Task-Subtask and Conversational Domain Dimensions of DATE